TCU-AM-15-112

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Leveraging Structural Characteristics of Interdependent Networks to Model Non-linear Cascading Characteristics

29 June 2015

Anita Raja

Albert Nerken School of Engineering

The Cooper Union

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

ABSTRACT

The overarching goal of our multi-year research agenda is to proactively model the non-linear cascading effects of interdependencies in Major Defense Acquisition Program (MDAP) networks. This document captures the progress that my team and I have made for the duration of this project. Specifically, we discuss the decision support architecture we describe our progress towards a scalable, automated approach for extracting and analyzing the data in the form of Selected Acquisition Reports (SAR) and Defense Acquisition Executive Summaries documents of a network of MDAPs to support a decision-theoretic risk prediction model. Automation is necessitated by the volume and complexity of the data. We will discuss the role of topic modeling, image extraction and identification of topological features of the MDAP network in this approach.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

The Joint Space of Major Defense Acquisition Programs Networks

It has been shown that data is the foundation for decision-making in the acquisition environment. The DOD has spent a significant amount of effort working across the organization to identify useful sources of data and conduct analysis. The importance of studying Major Defense Acquisition Programs (MDAP) interdependencies to acquisition research was emphasized during the 2012 NPS Acquisition Symposium by the introduction of a new panel titled Predicting Performance and Interdependencies in Complex Systems Development. Prior research has established that MDAPS are demonstrably interdependent, and that they can be thought of as networks of interdependent programs [Lewin 1999, Flowe et al., 2009]. Also, the acquisition paradigm established in statute 10 U.S.C. 2434, policy DoD 5000.02, and regulation tends to favor the notion of MDAPS as being independent, which would cause exogenous factors caused by interdependence to be overlooked or misinterpreted.

At the 2012 Acquisition Symposium, Dr. Frank Kendall III, the Under Secretary of Defense for Acquisition, Technology, and Logistics (USD[AT&L]), emphasized the role of achieving affordable programs that execute well and improving efficiency (via Better Buying Power and other initiatives) as key DOD strategic priorities. Historically MDAP performance is evaluated from an individual program point of view without emphasizing the dynamics of joint space. Our multi-year research investigation is based on the hypothesis that the poor performance of the MDAPs (various breach conditions) can be attributed to local as well as non-local sources that result due to interdependencies among the MDAPs. Therefore, it requires a network-centric, instead of an individual program-centric approach to capture the dynamics of the joint space.

In this report, we discuss our progress towards gaining a deeper understanding of interdependencies among Major Defense Acquisition Programs (MDAPs) by examining the various information sources including situational Selected Acquisition Reports (SARs), Defense Acquisition Executive Summaries (DAES) and Program Element documents (PE or R-docs). Our goal is to establish a statistically significant correlation between the state of MDAP network dependencies and their consequences.

In the first phase, we began a systemic investigation to support our network-centric hypothesis. We conducted a manual analysis of the MDAP performance reports that include DAES and SAR data belonging to a small network of MDAPs to determine the local and non-local issues that affect MDAP performance (Raja et al., 2012). In this work, we identified interdependency among programs resulted in cascading effects of issues across its neighbors. It was shown that reasoning about the success or failure from a network centric perspective would assist the programs managers to understand the consequence/risks of their decisions from a holistic point of view and hence would assist them to implement a strategic action to reduce the risk of failure. In addition to this, we recognized the need to analyze the data from the entire set of MDAPs in batch form to be able to build good decision models for “what-if” analysis.

The sheer volume and complexity of the data has led us to identify, as part of the second phase of this project, methods for automating the data extraction, network analysis and construction of the decision model (Raja et al., 2013). In this paper, we report and discuss the progress we have made in implementing and evaluating this automated approach. We have developed a decision-support framework would facilitate the cascading risk analysis. It is made up of multiple interacting modules including the **MDAP Network Identifier (MNI_MOD)** that represents the joint MDAP space in which MDAPs form funding interdependencies; the **Interdependency Index Determiner** (**IID_MOD**).

The significance of the research is three-fold:

- Aims to forge new ground in identifying the effects of interdependency on acquisition and, if needed, uncovering early indicators of interdependency risk so that appropriate governance oversight methods can then be isolated;
- Provides insight into automating the data extraction and analysis process by leveraging algorithms for decision support as well as image and text analysis.
- Examines the role of topological characteristics of interdependent MDAP networks to facilitate “what-if” analyses.

Research Methodology

To perform this study, we designed a methodology that involved developing the decision support architecture described in Figure 1. This architecture and associated control flow facilitate the study of cascading risk analysis from an individual MDAP perspective.

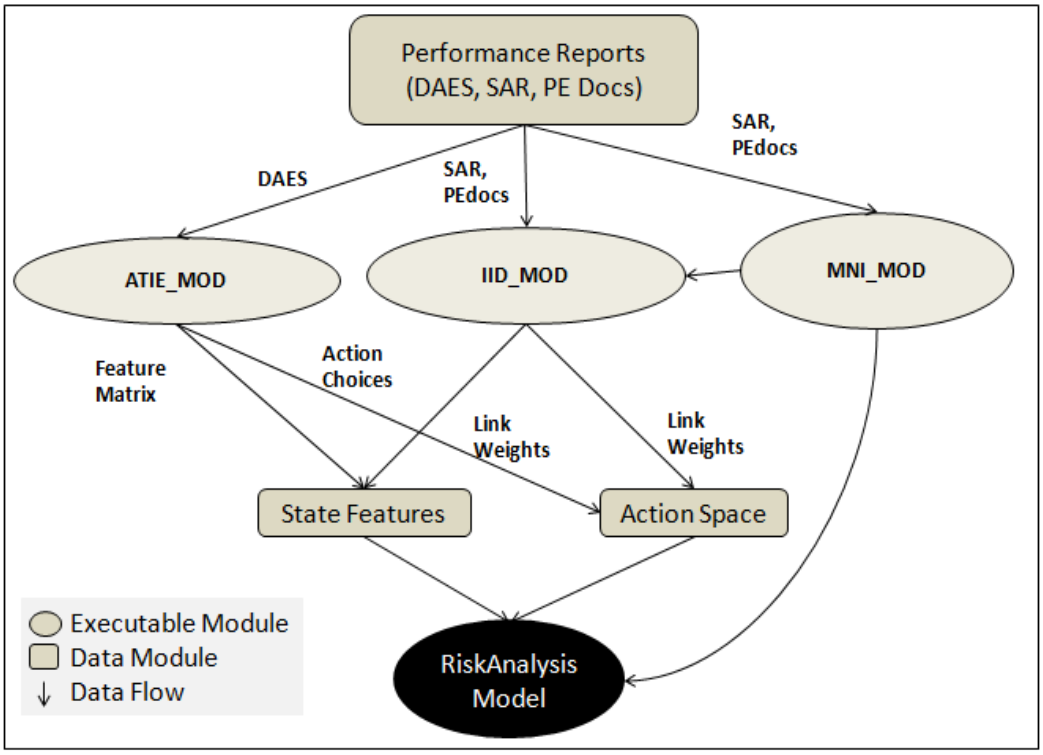

Figure 1: Decision Support Architecture for MDAP Cascading Risk Analysis

The main components are

- **Performance Reports Module:** It consists of the following performance reports: Monthly DAES reports that provide an early-warning report on the status of some program features such as cost, schedule, performance, funding etc.; SARs that summarize the latest estimates of cost, schedule, and technical status to be reported annually in conjunction with the President's budget ; Program Element (PE) documents (called PE docs or R-docs) that are used to justify congressional budgeting process.
- **Automated Text & Image Extraction Module (ATIE_MOD)** extracts key information from the performance reports to populate the state features as well as the action space of the decision model.
- **MDAP Network Identifier Module (MNI_MOD)** Based on these performance reports, the network neighborhood of the principal MDAP consisting of the MDAPs that share funding sources (PEs) is generated from an RDT&E perspective**.**
- **Interdependency Index Determiner** (**IID_MOD)** computes link weights between the interdependent MDAPs. These link weights are mapped to state features. Finally, the state features will be used to build a decision-theoretic model for "what-if" analyses.

Summary of Findings

Our findings indicate that the proposed decision support architecture would facilitate cascading risk analysis in MDAPs. We also confirmed that using state-of the-art extraction technologies including Latent Dirchlet Allocation (LDA) topic modeling algorithms as a basis, we have successfully developed automated text and image extraction techniques to extract features from various types of MDAP performance reports. In addition to this automated extractor module, we have developed two executable modules to identify the RDT&E funding network of MDAPs (MDAP Network Identifier) and to compute the weight of the links among the neighbor MDAPs (Interdependency Index Determiner). We tested and evaluated these algorithms on a small network of MDAPs and showed that the performance of the automatic extraction algorithms for Issues and Actions in the DAES reports was comparable to the performance of manual extraction

We now provide an in-depth discussion of the executable modules illustrated in Figure 1.

Automated Text & Image Extraction Module (ATIE_MOD)

The ATIE_MOD supports data extraction from raw data. The results of our previous work (Raja et al. 2012) while insightful and meaningful, involved tedious manual analysis of the defense reports. The reports consist of a voluminous combination of structured and unstructured data; this restricted the study to a small set of programs. With 'n' MDAP's and 12 reports a year, each MDAP accumulates n*12 reports where each report holds approximately a data of size 1MB. This estimation of data size multiplies with the number of years.

The performance of the MDAPs is reported periodically, once every month in a DAES' report in the form of Power Point files, Abode Acrobat files or as Word documents. A report includes textual and pictorial description of the health/performance of a MDAP. The textual information is descriptive of the ISSUES (problems or issues of concern) and the corresponding corrective ACTIONS considered for a program. For example, Table 1 shows the ISSUE SUMMARY page of a DAES' report which includes the ISSUE of 'unable to forecast cost' for a MDAP and necessary ACTION considered to resolve the ISSUE include 'reducing the headcount' of a team.

Images are used to represent the status of a program in a tabular matrix format, called the Program Status Matrix (PSM), for the past 3 months, current month and for future 8 months. Each row indicates the status with respect to cost, schedule, performance, funding and life cycle sustainment for Acquisition Program Baseline (APB)[1] and contract respectively. Green indicates that all contracts/APB requirements will be met, yellow indicates contracts and APB requirement that have problems but can be resolved and red indicates the contracts that will not be made available. PSM is always accommodated in the first of the report along with the details of the MDAP program name and the date of publication of the report as in Figure 2.

[1] Acquisition Program Baseline (APB): Baseline that reflects the threshold and objective values for the minimum number of cost, schedule, and performance attributes that describe the program over its life cycle.

Table 1: Example of ISSUE and ACTION summary for a MDAP

ISSUE SUMMARY	
ISSUES	ACTIONS
Cost Control- A Team, has been unable to provide an accurate forecast of projected cost. The Program Manager and [2]*(DOD office) are concerned that cost control problems will continue if costs are not accurately forecasted and closely monitored.	Continued engagement by the Program Manager and *(DOD office) *(person) is planned with *(team) management to address cost control issues. *(team) has detailed manpower projections and is measuring each organization's effectiveness in reducing headcount, a metric considered key to managing costs. *(team) headcount has been reduced, although at a rate to date slower than planned by the contractor, yet within the PM's estimate. With the recent loss of all remaining schedule margin, the PM's projected 6-month delay in * delivery seems highly probable, thereby increasing contract cost beyond the PM EAC.

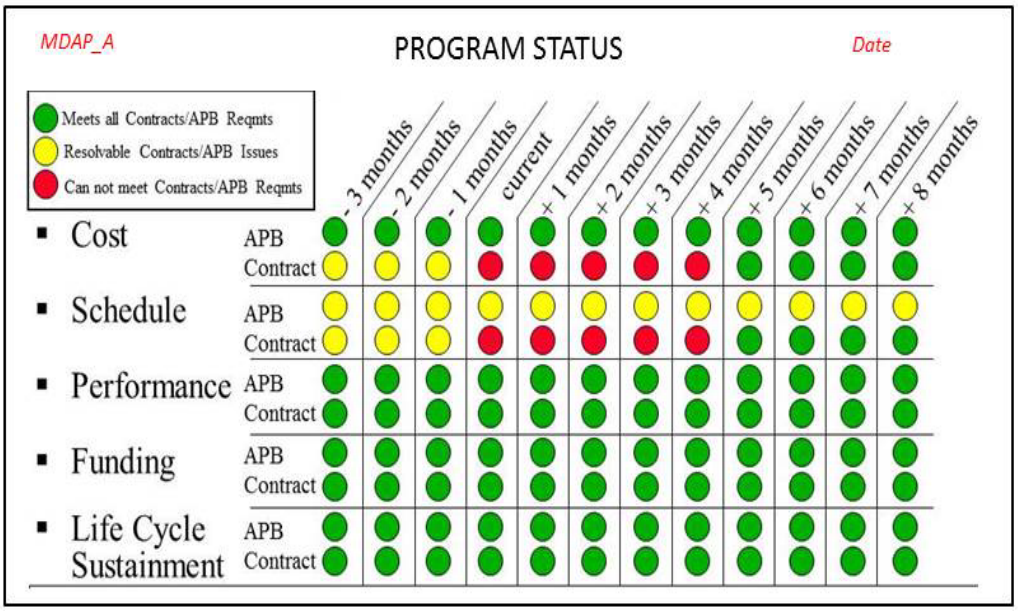

Figure 2: Example Program Status Matrix for a MDAP

We have developed a scalable and automated methodology that performs end-to-end image and text extraction and analysis on the DAES reports. This process will take as input as DAES reports of the program network and output the features that will populate the state and action space for the decision theoretic model. The text data is extracted using python scripting and analyzed using Latent Dirichlet Allocation (LDA) topic model (Blei, 2002) to discover the topics discussed in the reports. And the pixel data is extracted using java libraries and analyzed by the methodology of Topological structural analysis of images

[2] Throughout this document, in order to maintain confidentiality of the (FOUO) For Official Use Only Defense Reports data, we obfuscate program identifiable information using an *, where possible we mention the information type in brackets

The ATIE_MOD iteratively executes the following 5-step algorithm. The input to the algorithm consists of a batch of DAES reports belonging to a single MDAP.

ATIE Algorithm

Step 1: Prepare the data by converting DAES reports in Power Point or word format to a common Acrobat Adobe PDF standard format using 7-PDF open source software.

Step 2: Automate text and image extraction from DAESs reports using open source software: XPDF and ImageMagick. Python scripting is used to automate Step1 and 2 for batch processing of reports and for extraction of text from formatted text files generated by the XPDF toolX.

Step 3: Analysis of text using Probabilistic Topic Models (Blei, 2012), a machine learning algorithm to uncover the patterns of text in the DAES report. For example, the topics uncovered for the text in Table 1 will be "Unable to forecast cost" and "reduce headcount".

Step 4: Image analysis of the Program Status Matrix is done using Topological Structural Analysis of Images (Suzuki & Keiichi, 1985) to identify objects in an image and represent it in a numeric format with red circles represented with a value of -1, yellow as 0 and green as 1.

Step 5: Computation of Feature Matrix using the result of Text and Image analysis.

Step 6: Repeat steps 1 through 5 for all the MDAPs in the network.

We now elaborate each of these steps:

Data Preparation: Conversion of Power point/ word to pdf

DAES reports in Microsoft Power Point or word format are converted to a common PDF standard format and followed by extraction of text and images from pdf. This process is encapsulated in a python script called Master script to support batch processing of the reports. 7-PDF Maker (7-PDF) is an open source software which supports command line and GUI interface and Is used for the conversion to pdf format. Figure 3 captures this process.

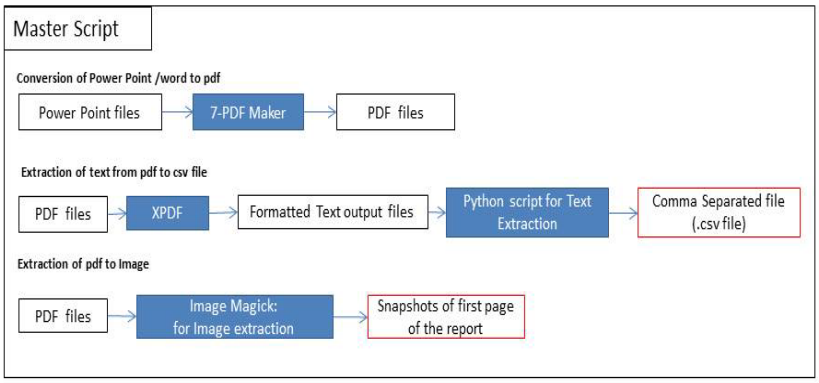

Figure 3: Parallel processes to prepare the data

Text and Image Data Extraction

Extraction of Text from PDF: PDF files are converted to text files using Xpdf converter, an open source software by Foolabs (XPDF). Command line execution of the software supports options for detailed conversion of report to text files. For example '–layout' option is useful to retain the layout of the pdf file. It is important to preserve the layout in order to identify the columns of the 'Issue summary' table. Hence conversion of pdf using Xpdf with layout option will generate formatted text files. Python scripting is used to extract text from formatted text file that includes the name of the program, Year, Month, list of ISSUES and ACTIONS. The ISSUES and ACTIONS are written to separate .csv files.

Extraction of Image from PDF: The Program Status Matrix (PSM) in the DAES report is a matrix of circular shapes supported by Microsoft Power Point or word. The PSM is captured in its completeness by conversion of the pdf page that contains the PSM to an image. A pdf can be converted to an image using ImageMagick (ImageMagick) that is open source software suite for to create, edit, compose, or convert bit map images.

Text Analysis using Topic Models

Topic modeling (Blei, 2012) is a machine-learning algorithm to identify topics in a document. The model works on the assumption that every document conveys information related to a set of topics. Latent Dirichlet Allocation (LDA) is a one such topic model (Blei et al., 2003). The model works on the 'bag of words' concept and it is efficient to discover patterns (groups) of words that appear in a document. These groups of words are tagged as abstract topics. In simple terms, when we read a document we highlight text with colors. If the same colored text is grouped together, each group forms a topic. The result of LDA is a probabilistic distribution of topics in a document and a topic is a probabilistic distribution over dictionary of words.

The LDA model parameters include:

1. **Observed variable**

   $\mathbf{W_{d,n}}$: The parameter represents the words in a document. It is the only Observed variable (known variable).

2. **Latent Variables**

   $Z_{d,n}$: The parameter represents the topic assignment value for a word.

   θ_d: The parameter represents distribution of topics for a document.

   β_k: The parameter represents distribution of dictionary of words across topic.

3. **Model Hyper parameters**

   α: is the parameter of the Dirichlet prior on the per-document topic distributions.

   η: is the parameter of the Dirichlet prior on the per-topic word distribution.

   The Hyper parameters α and η are the initial values for distribution of topics in a document and distribution of words in a topic respectively, that are set to a value greater than 0 to support iterative probabilistic computation.

Documents that serve as input to the model are broken down into words. We create a comprehensive dictionary of words by eliminating stop words that do not have a significant contribution to the meaning of the documents. These include 'the', 'these', 'a' and numeric and special characters like '@','1' and so on. Each document is represented as a vector of numeric values where a value in the vector corresponds to a word in the dictionary.

Table 2: Document Term Matrix

	Dictionary		
	Word 1	Word 2	Word 3
Document 1	12	10	1
Document 2	3	5	0
Document 3	9	8	29

Table 3: Assignment of topic to words in a document

Topics	1	4	2	1
Document 1	Word 1	Word 5	Word 3	Word 4

A numeric value that corresponds to a word can be represented with different meanings. For example, the frequency of occurrence of a word in a document can either be represented as, a tf-idf (term frequency- inverse document frequency) value or as a probabilistic distribution of topics for a word.

The Document Term Matrix (DTM) computed for documents as in Table 2, is the input to the model and Table 3 is a sample output:

The following are the steps of the LDA algorithm:

Step 1. Identify the number of topics (k) for input data set.

Step 2. Initialize every word in a document to randomly belong to a topic within k as shown in Table 4.

Table 4: Assignment of topic to words in a document.

Topics	1	4	2	1
Document 1	Word 1	Word 5	Word 3	Word 4

Step 3. Summarize the distribution of topics across the documents to understand the distribution of words across topics as in Table 5.

Table 5: Topic Distribution for a document

		Topic 1	Topic 2	Topic 3
Frequency of words in a topic	Word 1	1	2	1
	Word 2	9	7	5
	Word 3	10	8	6

4. Recompute the topic assignment for a word in a document based on its current numeric value in topic assigned and also based on the topic assignment of the neighboring words in a document. The model works iteratively to identify the set of words that could possibly occur together across the document and title them with a topic.

Equation 1 is the basis of the LDA model and is executed at every iteration. The variables in the equation that contributes to assignment of a topic to a word are as listed below,

1. The current topic assigned to the word, z.
2. The current distribution/occurrence of the word across topics, β.
3. The current distribution of topics over the document the word is present in, θ.

$$p(\beta_{1:k}, \theta_{1:D}, w_{1:D}) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d)$$
$$= (\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d.n}|\beta_{1:k}, z_{d,n})) \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Text Analysis Case Study

We now discuss the text extraction, analysis and evaluation for an MDAP called MDAP_A[3] as a prototypical case study. MDAP_A has 150 records of ISSUES and ACTIONS from DAES reports over a period of 6 years from **** through ****. Program MDAP_A is selected because it has a good repository of reports for most of the months over 5 years. For confidentiality reasons the program is referred to as MDAP_A.

a. Data Representation: The input data has to be preprocessed to eliminate noise in the data and create valuable numeric vectors for good performance of machine learning algorithms. This phase is commonly referred as, the data preparation phase. The input corpus is a set of unstructured documents. The reason for pre-processing of the data is to convert the unstructured data to a structured format where a document is represented as a numeric vector of values corresponding to the dictionary of words of the input corpus. The Stanford Text Mining Tool (STMT) creates the numeric representation of data with assistance in building the dictionary of the corpus. It is important to create a dictionary with significant words that will support to highlight the meaning of the documents. For example, words like play, ball, score are important in comparison to words like the, an, 12. Hence the process of creating a dictionary (Weiss et al., 2010) involves pre-processing of the documents using a set of filters including elimination of stop words, case conversion to lower case, and tokenization (a process to identify tokens/words in the documents using delimiters like space, new line character). The Term Minimum Document Filter is used to eliminate outliers that are most commonly or rarely used, as they do not highlight the meaning of the document. For example, misspelled words.

b. Parameter Estimation for the LDA model: This is the art of using a model for different applications demands fine-tuning of model parameters for appropriate results.
Number of topics (k) for input data: Quality of results of modeling can be measured using a factor called perplexity (Asuncion et al. 2009). Perplexity is a measure of model's ability to infer the topics in unseen documents. In simpler terms, it is a measure that defines how confused a model is. The perplexity script is created to support k-fold cross validation by splitting the input data as training and test data. Training data as the name suggests is used to train the model and test data is used to measure how surprised/perplexed the trained model is to the unseen data. Lower the perplexity value the model is more confident. Normal expected behavior of the model is that with the increase in the number of topics the perplexity reduces. We observed fluctuations in the perplexity measure and hence to normalize the measure, we repeated the process over five times and considered an average of the perplexity measure.

[3] the name of program MDAP_A is withheld since the "data is classified as official use only" FOUO

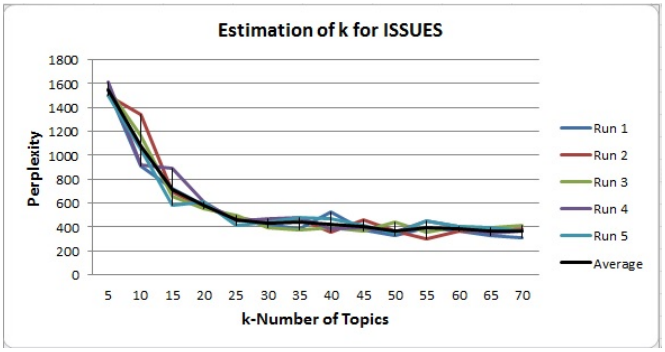

Figure 4: Number of topics for ISSUES estimated using perplexity measure

In Figure 4, we see that the value of perplexity stabilizes k equal to 20. The model is executed for a 'k' value of 20 and it is observed that the topics generated were sparse and upon re-investigation, comparatively meaningful results were obtained for 'k' value of 15. Hence the number of topics is set to 15 for the ISSUES input file with 150 records. A similar procedure is followed for ACTIONS input file as represented in Figure 5 and hence 'k' value of 10 is selected.

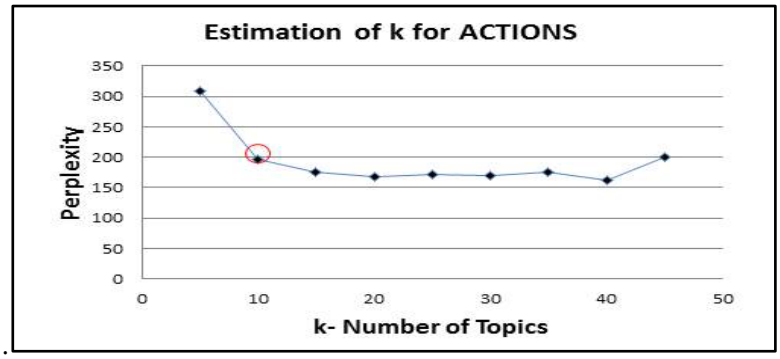

Figure 5: Number of topics for ACTIONS, estimated using perplexity measure

Dirichlet hyper parameters α and β: α and β are the initial values for distribution of topics in a document and distribution of words in a topic respectively which is a value greater than 0 to support iterative probabilistic computation. A smaller value of α and β will produce finer topics of specific interest (Griffiths & Steyvers, 2004). In this work, αand β values are defined to have lower values of 0.01.

Convergence of the model: Data convergence is the measure of the stability of the results of modeling that can be arrived at; by estimation of the number of iterations (N) the algorithm is to be executed (STMT). The results of modeling can be verified for stability by observing the change in the probability distribution values of words across topics. When the change is minimal the model is said to have converged. The factor 'N' is dependent on the size of the data. For a data size of 150 records 'N' is identified to be 800.

c. Modeling of data: The behavior of a model is dependent on the input data and the estimated model parameters. Therefore, a model has to be trained for an input domain and for the respective estimated parameter value to create an independent intelligent system. The trained system is then tested using different data from the same input domain.

Training Phase: It is typical in machine learning for a subset of the data, called training data, to be used to create Topic. The process is captured in Figure 6. DAES reports of MDAP_A for 5 years from **** through **** is the input data for training the model.

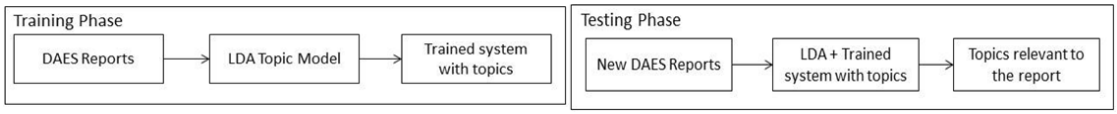

Figure 6: Process to capture the training & testing phase

Testing Phase: The trained model is tested on unseen data to evaluate its performance as shown in Figure 7. Results of testing will include the topics relevant to the report and a list of words for each topic that can be verified with the corresponding topic of the trained system.

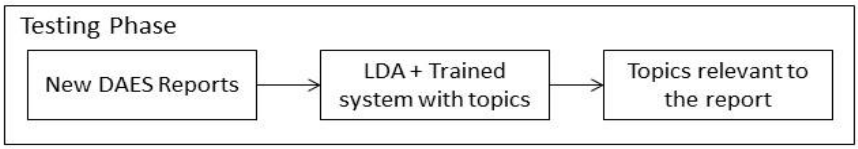

Figure 7: Block diagram to capture testing phase

The text extraction process results in the creation of the ISSUES and ACTIONS .csv file that contains the ISSUES and ACTIONS recorded in the DAES reports respectively and is analyzed using the LDA topic modeling module of Stanford Topic Modeling Tool (STMT) kit. A Collapsed Variational Bayesian sampling version (The, Newman, Welling, 2007) of LDA is used.

The result of Topic Modeling is a probabilistic distribution of latent topics in the data and the topics are a probabilistic distribution over dictionary of words. The results include abstract topics labeled as 'Topic 0', 'Topic 1' and so on. A topic is represented by list of words that are sorted in descending order of their probability distribution values. Hence in order to understand a topic, the top few words listed for a topic should be formalized together to conclude on a label that is descriptive of the words. For example, as in Figure 5, 'Topic 0' with words 'tax, money, income, paid, pay, trust' can be associated to the label 'Tax'. We plan to use classification algorithms for this labeling problem in future work.

Image Analysis

Analysis of images involves exploring the pixel information of the image to identify the objects and its descriptions like size, color or orientation. The OpenCV (Open source Computer Vision Library) (OpenCV) is an open source computer vision and machine learning software library that offers implementations of image processing algorithms that are compiled in the 'Imgproc' package. The 'findContours' method included in the imgproc package is an implementation of the Contour identification algorithm (Suzuki & Keiichi, 1985) which is used to identify the structure of the binary image and hence to explore the contour/objects in the binary image. The parameters to identify the contours using the 'findContour' method are as listed below,

- Input Image in binary form: An image is converted to its binary from using 'Canny' method supplied in the 'imgproc' package.
- List to hold the set of points for the contour identified.
- A temporary image used by the algorithm to update the intermediate results of processing of the image.
- Mode: Contour retrieval mode that defines that structure of hierarchy of the image to be preserved.
- Method: Contour approximation method that defines the points of a contour to be stored to the list of identified contours.

The forecast of the status of MDAPs is captured in the first page of DAES report as program status matrix (PSM) using colored circles. The green indicates the contracts that are met, the yellow indicates resolvable contracts and the red is for contracts that cannot be met.

As described earlier, the first page of the DAES reports is converted using ImageMagick. These images are analyzed to identify the colored circular shapes and represent them in wise numeric format with value of -1 for red circles, 0 for yellow circles and 1 for the green circles as represented in Table 2. Image analysis involves scanning of the pixel data to extract information of the image that include the shape and color of the objects in the image. For this purpose, OpenCV (Open source Computer Vision Library) (OpenCV) java libraries is used in this work. OpenCV is an open source java library that is extensively used in the field of computer vision for images analysis.
Image analysis involves the following steps,

1. **Preprocessing of the Image:** The colored image is converted to a binary image/ grayscale and the edges of the items in the image are highlighted. 'Imgproc' (OpenCV Java) package of OpenCV supports the methods cvtColor() and Canny() to convert a colored image to binary image and highlight the edges respectively.

2. **Contour identification algorithm to find contours/objects:** Highlighted edges of the binary image is used identify contours in the image using the findContours() method of Imgproc package (OpenCV Java) of OpenCV library. findContours() method returns a list of array of points where an array corresponds to the points of a contour. The size of the list of contours is large as it includes all the contours in the image that is the circles, lines and characters of the text. For each contour, the list of points is traversed and a compact circle that can connect the points of the contour to form a circle is identified by minEnclosingCircle() method of ImgProc package. The minEnclosingCircle() method returns the center and radius for the circles identified.
3. **Process the circles to eliminate noise:** The results of minEnclosingCircle() function include enormous number of circles because the method identifies the minimum possible circle around all the objects including characters like 'a', 'm','n,'o'. Also, for a few circles in the PSM inner and outer circles are identified and hence there are overlapping circles in the PSM.

   To eliminate the noise and obtain the circles of interest, the results of minEnclosingCircle() is processed by using conditional filters and stored in a TreeMap data structure supported by java TreeMap package (Java Libraries). The TreeMap data structure holds a <KEY, VALUE> pair in ascending sorted order of the keys by default behaviour. The Y coordinate of the center of circle is the KEY and the list of the X coordinate of the center of the circle forms the VALUE.

The filters to process the circles include Radius of the circle, Pixel color of the center of the circle, Avoid duplicates in the TreeMap. This latter filter is used to avoid the overlapping circles in the PSM amongst the colored circles. A pixel distance check between the center of the circle of focus and the center of existing circles in the TreeMap, is measured before a circle is added to the TreeMap. If the circle of focus is not in the predefined pixel distance threshold of ±10, it is added to the TreeMap.

4. **Extract information of the circles:** The processed circles in the TreeMap are scanned to eliminate rows with one circle. The resultant set of circles is transformed to numeric format and printed on to a .CSV file.

Feature Computation

The result of text analysis is a compilation of the labeled trained and tested DAES reports for the ISSUES and ACTIONS of a MDAP while the result of image analysis is a comprehensive representation of the PSM for all the DAES reports of a MDAP. Hence a MDAP is represented in its completeness in a .csv file with rows representative of features and columns representative of the DAES reports.

MDAP Network Identifier Module (MNI_MOD)

The following steps facilitate the creation of the RDT&E funding network for MDAP_X:

1. First, the RDT&E PE account is retrieved from the "Track to Budget" page of the respective SAR file of a particular year. The PE account that funded MDAP_X for RDT&E is PE_abc.
2. The above mentioned PE document is downloaded from the RDDS website.
3. The first page of the PE document provides a list of the programs that are funding recipients of this PE account. These programs include both MDAP and non-MDAPs. As an illustration, Table 6 shows a list of funding partners that received RDT&E funds via PE_abc:

Table 6: Funding table for FY 2007 from PE_abc (from 2009 document)

Programs	Cost ($ in millions) for FY 2007
Non-MDAP_a	77.676
Non-MDAP_b	2.075
MDAP_X	**645.851**
Non-MDAP_c	0.0
Non-MDAP_d	2.876
Total PE Cost	728.480

Among these programs, only MDAP_X is an MDAP. The PE document reveals that Non-MDAP_a channels a large fraction of its received fund to an MDAP named MDAP_Y. Therefore, it appears that MDAP_X and MDAP_Y are the two MDAPs that share same PE account. This is also confirmed by looking at the "Track to Budget" page of the SARs of these two MDAPs. Hence, MDAP_X and MDAP_Y are the RDT&E funding partners for that year.

4. Step 1-3 is performed for the successive years of interest and an evolving MDAP_X RDT&E funding network is created.

Interdependency Determiner Module (IID_MOD)

This module computes the link weights among the interdependent MDAPs. A **link** is defined as a relationship between the PE account and the respective RDT&E fund receiving MDAPs. Therefore, the **link weights** for the MDAPs are the percentage of the funding that each MDAP receives from their *respective PE account*. The link weights for the funding partners are computed for each year using the following 3 step process:

i. The "Accomplishments/Planned Programs" section of each program that is listed in the first page of the respective PE document provides the breakdown of funding. The exact received funding amount for the MDAPs is retrieved from these sections.

ii. The received funding is matched with the funding amount from the SARs of the respective MDAPs. This SAR funding amount is found from the page that records the "Annual Funding by Appropriation".

iii. For each MDAP, the received funding amount (retrieved from PE, as in step *i*) is divided by the total PE funding and the percentage of received funding is computed.

Empirical Evaluation

The objective of comparative study is to ensure the topics identified by automated tool analysis are comparable to those extracted by human experts.

Evaluation of automated analysis for ISSUES data

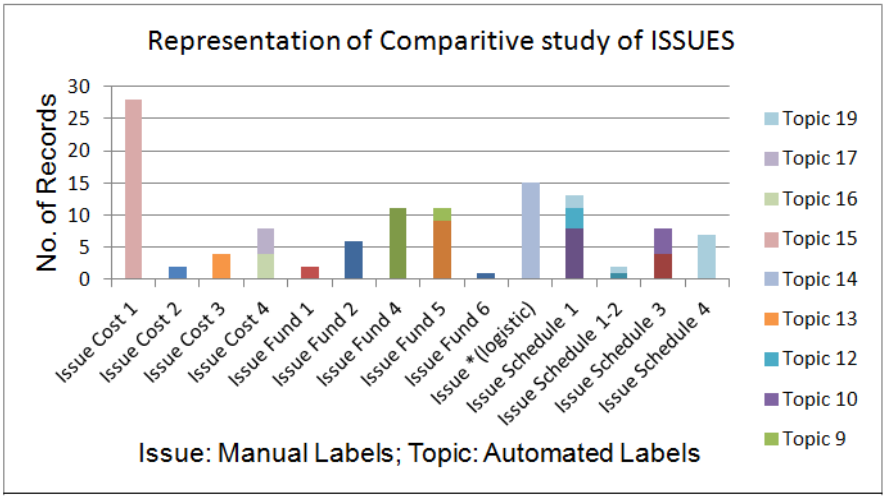

Figure 8: Comparative study of Manual analysis vs. Automated tool analysis for ISSUES data

In this comparative study, ISSUES and ACTIONS data from DAES reports of the program MDAP_A for a time frame of 5 years from **** through **** are analyzed.

For Manual analysis, the input data is categorized as main topics with sub topics if required and the input records are tagged to belong to one of the topics. The main topics are Cost, Schedule, Funding and Logistics.

For Automated tool analysis, the same input is analyzed using the topic model that provides a distribution of topics for the input records. Based on the sorted list of distribution values of topics, a record is associated with topics of higher probability values of 0.5 and greater. A record may be associated to one or more topics and this information is mapped to the input records. This process prompted us to differentiate records as simple and complex. Records associated to one topic are considered to be simple and more than one topic is considered to be complex. We first evaluate the effectiveness of the automated analysis by comparing to manual analysis by human experts on the same data. Figure 8 shows this comparison for ISSUES related to MDAP data while Figure 12 compares ACTIONS related to MDAP data.

In Figure 8, X axis includes ‘Issue: Manual Labels’, are the categories of manual analysis. For example ‘Issue Cost 1’; ‘Topic: Automated labels’ are the topics generated by the tool. For example ‘Topic 5’.Y axis is the number of records in each dimension. Analysis of the above graphic can be considered across four criteria.

Criteria 1: Exact mapping of abstract topics to manual issue type.
There are 14 manual issue types assigned for input data. The above representation indicates there exists one to one mapping amongst six manual issue type and topic. This observation is captured in the Table 7.

Table 7: Exact mapping of abstract topics to manual issue type

	Issues Type	Topic Number
1	Issue Cost 1- inability to accurately forecast cost	28
2	Issue Cost 2- Overrun costs due to technical delays	7
3	Issue Cost 3- Modifications to total cost to avoid going over budget	13
4	Issue Fund 1- Funding has been aligned with estimate	8
5	Issue Fund 4- Funding is being reassessed due to errors or changes in need	11
6	Issue *(logistic)- *(logistic) is unavailable for installation.	14

Criteria 2: An issue type mapped to more than one topic.
A few issue types are mapped to more than one topic. On close investigation of the data for these records reveal that the model is being more specific in further categorizing the issue type. Table 8 captures an explanation for this category.

Table 8: An issue type mapped to more than one topic

	Issue type	Multiple topic for an issue type
7	Issue Cost 4-Cost Overrun	Issue type is split across topic 16 and 17.The topics focus on different reasons for cost overrun.
8	Issue Fund 5 – Funding is needed because there is currently none	Issue type is split across Topic 6 and 9. The topics focus on different reasons, one, funding for launch of equipment and other for interface building.
9	Issue Schedule 3-International Collaboration Delay	Issue type is split across Topic 1 and 10. The topics focus on different reasons for delay, one, waiting for contract approval from another vendor and other is vendor needs to make initial set up and also approve contract to start the project.

Criteria 3: **Combine manual issue types to one topic**
Model identifies records of 'Issue fund 2 and fund 6 type' to be combined to a one category i.e. topic 0.Issue 'fund type 6' is of records in which the project funds are moved temporarily to another project and hence expecting a temporary fund shortfall. And 'fund type 2' is for records with shortfall of funds. So the tool pulls the two reasons together under one topic of fund shortfall.

Criteria 4: Complex records - Records with high probability to belong to multiple topics
An example of a complex record, Issue Schedule 1 is studied to understand the source of the complexity. The hardware related ISSUE in these records can be classified into 3 overlapping buckets,

- The product build was a failure and hence rework is required.
- The product build was a failure and hence rework is required and also the project was on hold expecting for a component.
- Project was on hold expecting for a component.

The overlapping reasons identified as different topics lead to the complexity of a record. Below, we capture the complexity graphically.

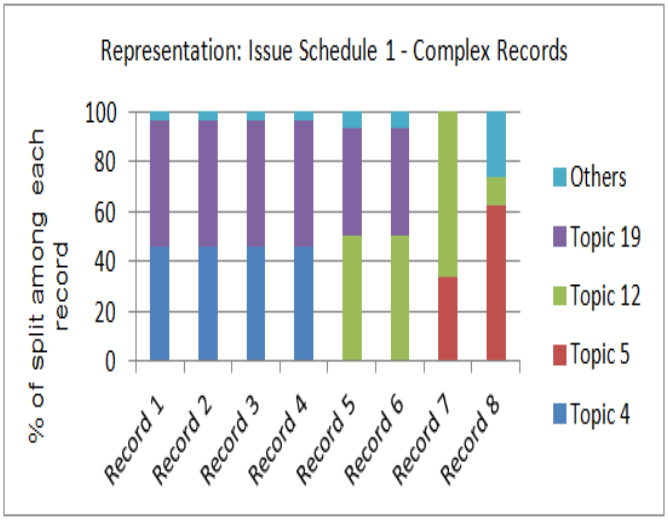

Figure 9: Complex records, records with probability to belong to different topics.

Evaluation of automated analysis for ACTIONS data

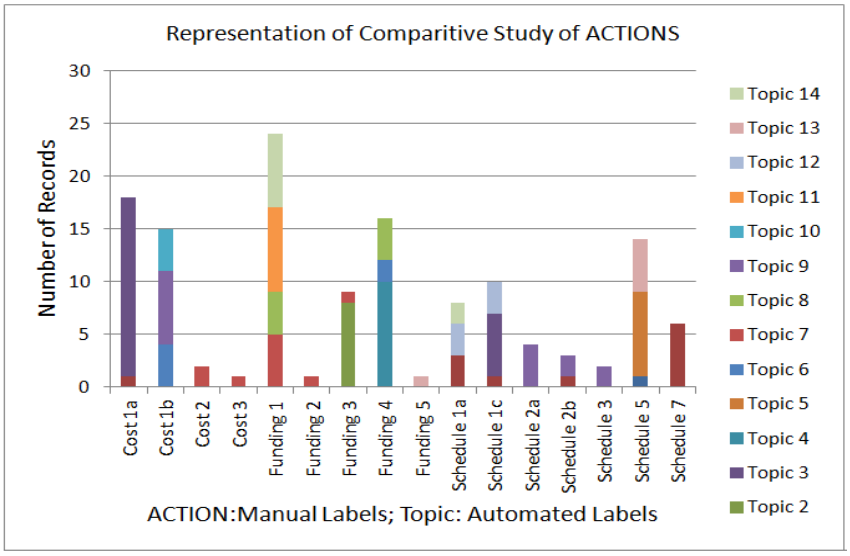

Figure 10: Comparative study of Manual analysis vs. automated tool analysis for ACTIONS data.

In Figure 10, X axis includes 'ACTION: Manual Labels', are the categories of manual analysis. For example 'Cost 1a'; 'Topic: Automated labels' are the topics generated by the tool. For example, 'Topic 0'. The Y axis is the number of records in each category.

Analysis of the above graphical representation can be spread across four criteria:

Criteria 1: Exact mapping of abstract topics to manual action type.

There are 16 manual Action types assigned for input data. The above representation indicates there exists one to one mapping amongst nine manual ACTION type and topics. This observation is captured in the Table 9.

Table 9: Exact mapping of abstract topics to manual ACTION type

ACTION Type	Topic Number
Cost 1a: Contractor Renegotiation	3
Cost 2: Monitor performance	7
Cost 3: Manage cost by reducing head count	7
Funding 2: Work with OSD and senior Leadership	7
Funding 3: Fund DoD *	2
Funding 5: Work with *(defense department)	14
Schedule 2a: Projection of schedule based on hardware and I&T progress	9
Schedule 3: APB baseline change	9
Schedule 7: Accelerate schedule for Early contract reward	1

Criteria 2: An ACTION type mapped to more than one topic.

A few ACTION types are mapped to more than one topic (Table 10). On close investigation of the data for these records reveal that the model is being more specific in further categorizing the ACTION type.

Table 10: An ACTION type mapped to more than one topic

	ACTION type	Multiple topic for an issue type
1	Schedule 5: Investigate alternate *(logistic)	Reasons for investigation are -There is an long delay in approval of *(logistic) by a foreign *(entity) and hence they investigate on alternate site -Initial Effort is initiated get approval to work on a *(logistic) and alternate *(logistic) are investigated as a backup
2	Funding 4: Monitor program to determine when to fund launch of *(logistic)	The sources of funding to launch the *(logistic) are two different authorities/groups.

C**riteria 3: Combine manual ACTION types to one topic**

- Model identifies records of 'Cost 2, Cost 3 and funding 2 type' to be combined to one category i.e. topic 7. ACTION 'Cost 2 and Cost 3' is from records in which the cost of the programs is monitored by tracking performance and reducing team head count. 'funding2' is of records which have a funding shortfall and hence the cost of the program is monitored. So the tool combines the two reasons together under one topic of Monitor Cost.
- Another instance in this category is the records 'Schedule 2a, 2b and 3'. 'Schedule 2a and 2b' are the actions considered for schedule delay that includes projection of schedule and updating of schedule respectively. 'Funding3' address the issue of shortage of funds by updating the schedule. The main course of action is "update of schedule" which is put together by the tool.

To summarize the comparative study, we observe that the results of automated tool analysis can be mapped to manual analysis and is advantageous in terms of the category 2, 3 and 4 as explained above, where the tool is precise and more specific to the ISSUE and ACTIONS associated with a record. Hence performance of topic modeling tool is comparable to human efforts. The comparative study has also helped use further understand the domain knowledge that we believe is important for analysis of results in our further research.

Evaluation of Text Analysis

The result of topic modeling of DAES reports includes the distribution of topics for the ISSUES and ACTIONS where the topics are represented with list of co-occurring words as represented in the Figure 11 and 12.

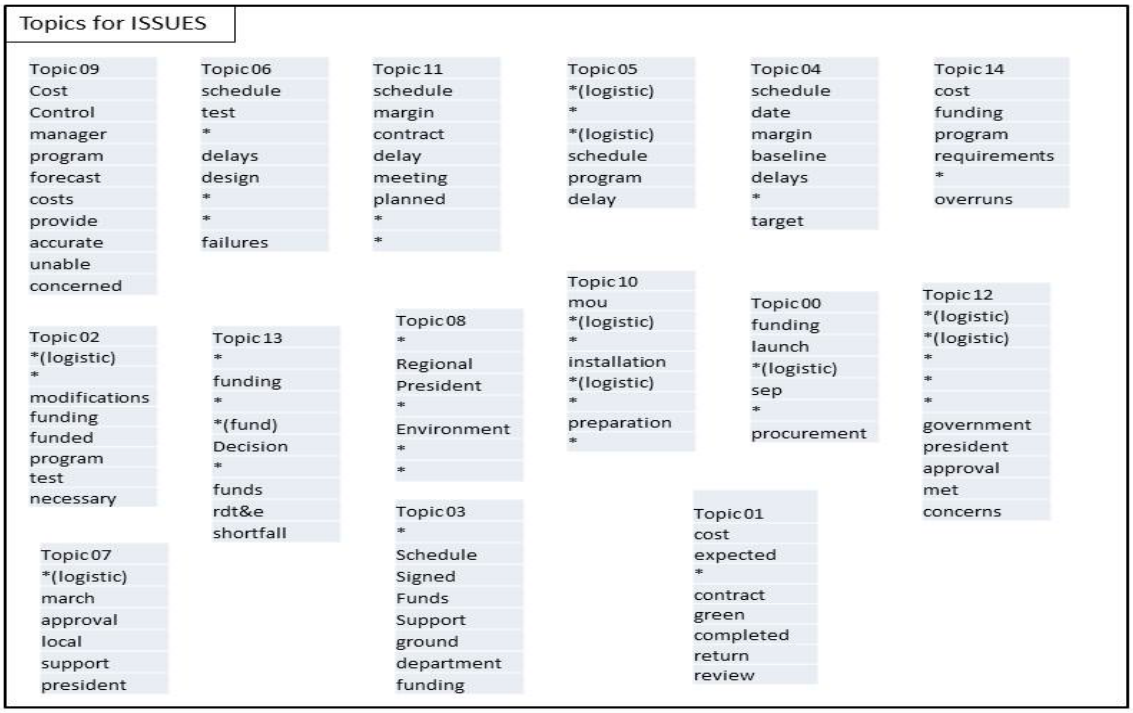

Figure 11: Topics for ISSUES

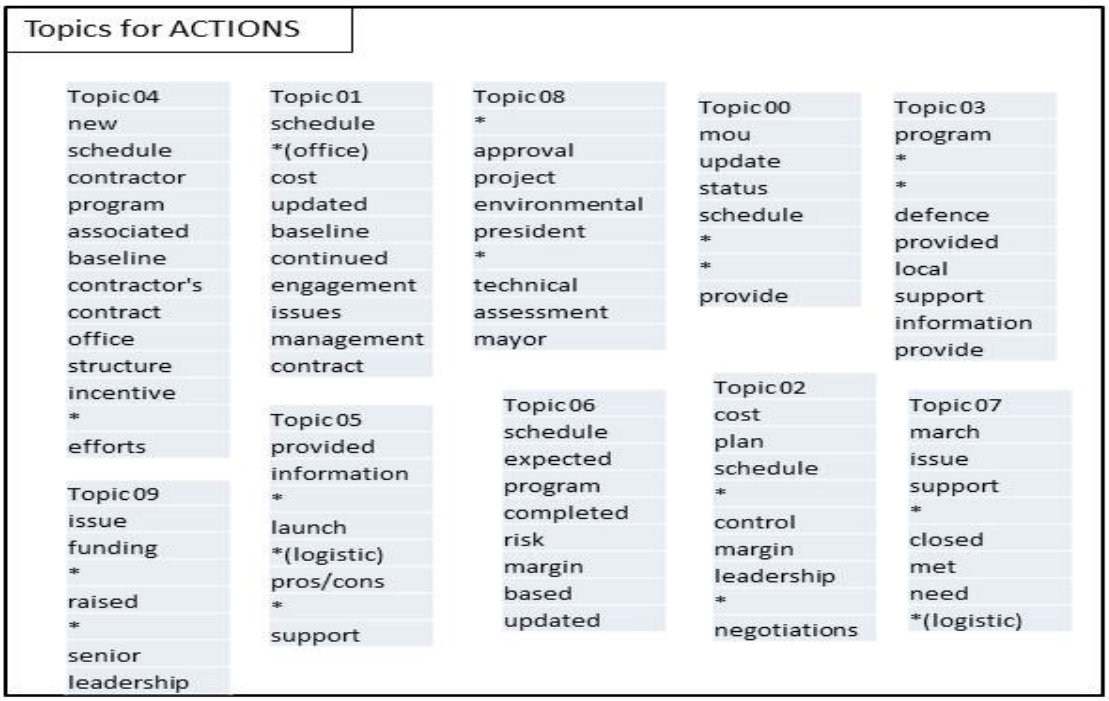

Figure 12: Topics for ACTIONS

Topics generated by the topic model are informative of the content in the reports. For example *'Topic 09'* for the ISSUES conveys the fact of *'unable to forecast cost'*. While most of the topics are informative a few are difficult of understand. For example, *'Topic 07'* in actions consists of words like *'march, issue, support, closed'* which is difficult to interpret.

The trained model is evaluated for an unseen data which are the reports of the year ****.
Evaluation of the ISSUES trained model:

Figure 13 shows the distribution of unseen ISSUES data across the topics of the trained ISSUE model and the Table 11 represents a sample of mapping of ISSUES to respective topics.

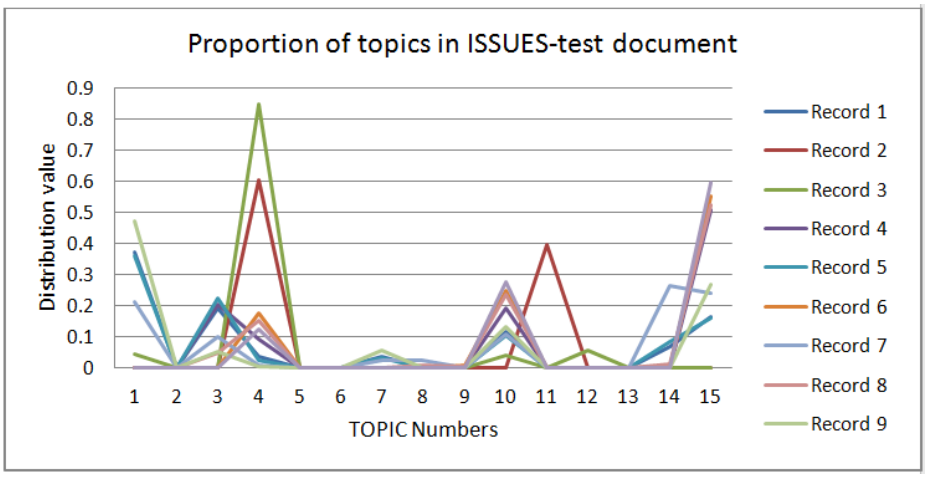

Figure 13: Proportion of topics in ISSUE testing data.

Table 11: Sample of results of evaluation of ISSUE trained model

Record	ISSUE	Topic	Topic Descriptors From Figure 16
Record 3	Requires funding to maintain *(logistic) for * launch	3	Schedule,funds, support, *(logistic),department
Record 7	Cost increase – unavailable on time funds	14	Cost,funding,requirement,overrun
Record 2	Schedule delay- delay in contract agreement	3,10	Schedule,funds, support, *(logistic),department, installation,equipment

Figure 14 shows the distribution of unseen ACTIONS-data across the topics of the trained ACTIONS model and the Table 12 represents a sample of mapping of ACTIONS to respective topics.

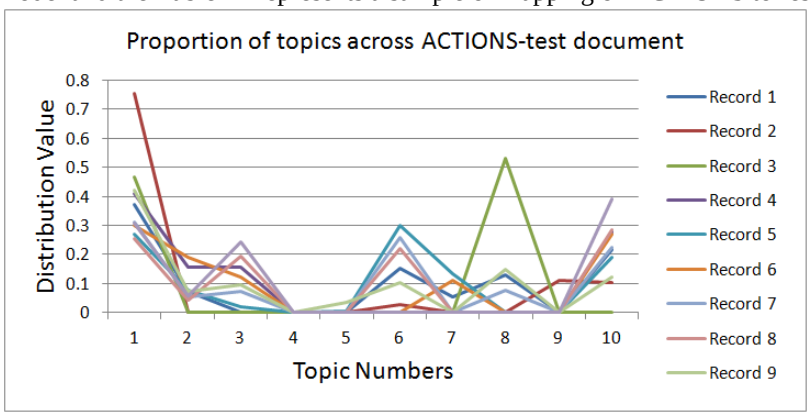

Figure 14: Proportion of topics in ACTION testing data.

Table 12: Sample of results of evaluation of ACTION trained model.

Record	ACTIONS	Topic	Topic Descriptors From Figure 14
Record 2	Schedule : Request for approval, status will be revisited	0	Update, status, schedule
Record 3	Fund: Prepare impact statement, status will be revisited	7,0	Update, status, schedule, support, closed, met, need
Record 10	Update requirement of fund, status	9,0	Update, status, schedule, issue, funding, raised

Evaluation of Other Modules

The RDT&E funding network that we have constructed based on the approach in the module description above includes two MDAP neighbors: MDAP_A and MDAP_B. This network spans across 2005-2012 and remains static during this time period.

Table 13 shows the funding table retrieved from the first page of the PE_xyz document that has funded MDAP_A. In order to find the breakdown of the received funding amount for MDAP_A and its partners, we looked into the "Accomplishments/Planned Programs" sections for MDAP_A and Non-MDAP_1. Table 14 shows the received funding from both PE account perspective and from the respective SAR files. This is done in order to match the funding Figures across these two types of performance reports. Table 13 indicates that the received funding for MDAP_A and MDAP_B as reported in the PE documents and SAR files are mostly matching. Moreover, the last two columns of this table records the link weights of these two MDAPs with respect to the common PE account.

Table 13: PE_xyz Funding

Year	**Total (PE)**	**NON MDAP_1**	**NON MDAP_2**	**NON MDAP_3**	**NON MDAP_ 4**	**NON MDAP_5**	**Congressiona l Adds**	**MDAP_ A**
Y1	155.3	56.7	0.5	12	0	1.7	-	84.4
Y2	447.061	47.914	0.683	17.567	-	-	5.688	375.209
Y3	527.432	51.792	1.189	19.227	-	-	5.757	449.467
Y4	728.48	77.678	2.075	0	-	-	2.876	645.851
Y5	715.169	98.371	11.846	7.472	-	-	4.055	593.425
Y6	625.193	113.931	10.28	0.212	-	-	0.797	499.973
Y7	481.831	79.956	1.049	2.509	-	-	-	398.317
Y8	410.015	18.026	0.607	-	-	-	-	391.382
Y9	258.811	17.476	4.155	-	-	-	-	237.18

Table 14: Comparison between PE and SAR funding for MDAP_A and MDAP_B; Link Weight Computation for MDAP_A & MDAP_B

Year	Total (PE)	MDAP_B (PE)	MDAP_A(PE)	MDAP_B (SAR)	MDAP_A (SAR)	Link Weight – MDAP_A	Link Weight – MDAP_B
Y1	155.3	55.389	84.4	64.1	84.4	54.34642627	35.66581
Y2	447.061	46.514	375.209	58.1	375.2	83.92792035	10.4044
Y3	527.432	51.341	449.467	53.5	449.5	85.21799967	9.734146
Y4	728.48	77.678	637.112	77.7	637.1	87.45772018	10.66302
Y5	715.169	88.204	591.438	87.7	591.3	82.69905435	12.33331
Y6	625.193	109.317	497.028	109.3	497	79.49993042	17.48532
Y7	481.831	78.956	369.243	79	369.2	76.63330089	16.38666
Y8	410.015	18.026	391.382	18	391.4	95.45553211	4.396425
Y9	258.811	17.476	237.18	17.5	237.2	91.64216359	6.752418

Conclusions and Future Work

Our overall goal is to study the RDT&E funding dynamics across MDAPs with the intent of uncovering early indicators of interdependency risk so appropriate governance oversight methods can then be isolated. In this paper, we discuss the design and evaluation of automated text and image extraction techniques we have developed to extract features from various types of MDAP performance reports. In addition, we have developed two executable modules to identify RDT&E funding network of MDAPs (MNI_MOD) and to compute the weight of the links among the neighbor MDAPs (IID_MOD). In our continuing work, the output of these three modules would contribute to the state features and action space of the cascading risk-analysis model. As part of our future work, we also plan to address the topic labeling problem.

Acknowledgement

I am grateful to Shalini Rajanna, Mohammad Hasan and Jagan Vujjini for their diligent work that ensured the success of this project. Shalini was a key contributor to the automatic information extraction algorithms while Jagan helped scale up the scripts to other MDAPs and organize the software packages. Hasan helped coordinate various aspects of the project and worked on the network analysis. I am thankful to Dr. Ansaf Salleb-Aouissi for her significant contribution towards extending existing image and text extraction algorithms to fit our problem. I am grateful as always to Dr. Maureen Brown for providing us access to MDAP data and for her deep insights into the world of MDAPs . I also thank Mr. Robert Flowe for his valuable input and time in helping us understand the data and related challenges from OSD's perspective.

References

Asuncion, A., Welling, M., Smyth, P., & Teh, P. Y. (2009). On Smoothing and inference for Topic Models, *Proceeding of Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Pages 27-34.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), Latent dirichlet allocation.. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M. (2012). Probabilistic Topic Models, *Communication of the ACM*, New York USA, Volume 55 Issue 4, Pages 77-84.

Cunningham, P., & Delany, S. J. (2007). k- Nearest Neighbour Classifiers, University College Dublin, Dublin Institute of Technology. *Technical Report UCD-CSI-2007-4*.

Eclipse IDE for Java Developers. Version, Kepler Service Release 1. *http://www.eclipse.org/*

Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences* **101** (Suppl. 1): pp 5228-5235.

ImageMagick software by ImageMagick Studio. Free software suite to create, edit, compose, or convert bitmap images. Version 6.8, October 2013, *www.imagemagick.org*

Java Libraries, Java ™ Platform Standard edition 6. *http://docs.oracle.com/javase/6/docs/api/java/util/TreeMap.html*

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification, *Proceedings of the 2007 conference on Emerging Artificial Intelligence, Applications in Computer Engineering: Real world AI systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, Pages 3-24.

Lewin, Y. "Application of Complexity Theory to Organization Science." *Organization Science*10:3(215-236). 1999.

Flowe, R. M., Brown, M. M. and Hardin, P. L. (2009). Programmatic Complexity and Interdependence: Emerging Insights and Predictive Indicators of Development Resource Demand. Technical Report prepared for Naval Postgraduate School.

OpenCV-Open source Computer Vision Library by iteseez, an open software for computer vision and machine learning algorithms, Version 2.4.6, July 2013, http://opencv.org/about.html

OpenCV Java Libraries, package for Image processing, Version 2.4.6, 2013. *http://docs.opencv.org/java/*

Raja, A., Hasan, M. R., & Brown, M. M. (2012). *Facilitating Decision Choices with Cascading Consequences in Interdependent Networks.* Report prepared for Naval Postgraduate School.

Raja, A., Hasan, M. R., Rajanna, S., & Salleb-Aoussi, A. (2013). *Leveraging Structural Characteristics of Interdependent Networks to Model Non-linear Cascading Risks.* Report prepared for Naval Postgraduate School.

R-tm, Text mining package, open source framework for text mining applications within R. Version 0.5-9.1, July 2013.*http://cran.rproject.org/web/packages/tm/index.html*

R-FNN, Fast Nearest Neighbor Search algorithms and Applications in R. Version 1.1, July 2013. *http://cran.r-project.org/web/packages/FNN/FNN.pdf*

Stanford Topic Modeling Toolbox by Daniel Ramage and Evan Rosen of *Stanford Natural Language processing Group*, Version tmt-0.4.0, September 2009.

Suzuki, S. & Keiichi, B. (1985). Topological Structural Analysis of digitized binary images by border following, *Journal of Computer Vision, Graphics, and Image Processing*, Volume 30, Issue1, Pages 32-46.

The, Y. W., Newman, D. & Welling, M. (2007), A Collapsed Variational Bayesian Inference algorithm for Latent Dirichlet Allocation, *Proceedings of Advances in Neural Information Processing Systems*, 19, 2007

Weiss, S. M., Indurkhya, N. & Zhang, T. (2010). Fundamental of Predictive Text Mining, Published by Springer ISBN 978-1-84996-225-4.

XPDF software by Foolabs . Open source viewer for Portable Document Format (PDF) files, Version 3.03, August 2011,*www.foolabs.com/xpdf*

7-PDF Maker software by 7-PDF company, Germany. Freeware PDF converter for creating PDF files, Version 1.4.1, 2013. *www.7-pdf.com*.

Biographical Information

Anita Raja is a Professor of Computer Science at The Cooper Union since August 2014. She was an Associate Professor of Software and Information Systems at the University of North Carolina in Charlotte for the duration of this project. She received her PhD in Computer Science from the University of Massachusetts Amherst in 1998 and 2003 respectively. Professor Raja's research focus is in the field of artificial intelligence, specifically as it relates to the study of decentralized control and reasoning in software agent systems operating in the context of uncertainty and limited computational resources.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net